

Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners

Year: 2023 | Citations: 295 | Authors: Allen Z. Ren, Anushri Dixit, Alexandra Bodrova

Abstract

Large language models (LLMs) exhibit a wide range of promising capabilities -- from step-by-step planning to commonsense reasoning -- that may provide utility for robots, but remain prone to confidently hallucinated predictions. In this work, we present KnowNo, which is a framework for measuring and aligning the uncertainty of LLM-based planners such that they know when they don't know and ask for help when needed. KnowNo builds on the theory of conformal prediction to provide statistical guarantees on task completion while minimizing human help in complex multi-step planning settings. Experiments across a variety of simulated and real robot setups that involve tasks with different modes of ambiguity (e.g., from spatial to numeric uncertainties, from human preferences to Winograd schemas) show that KnowNo performs favorably over modern baselines (which may involve ensembles or extensive prompt tuning) in terms of improving efficiency and autonomy, while providing formal assurances. KnowNo can be used with LLMs out of the box without model-finetuning, and suggests a promising lightweight approach to modeling uncertainty that can complement and scale with the growing capabilities of foundation models. Website: <https://robot-help.github.io>